利用游戏 log-file 预测学生推理能力和数学成绩——机器学习的应用*

孙　鑫 (1)；黎　坚 (1,2)；符植煜 (1)

([1] 北京师范大学心理学部; [2] 应用实验心理北京市重点实验室，北京　100875)

摘　要　以 360 名初中生为被试，使用推箱子游戏，结合游戏日志文件(log-file)和机器学习技术预测学生的推理能力和数学成绩。预测变量是从推箱子的过程数据中提取的一系列特征指标，结果变量是瑞文推理测验成绩和数学成绩，且均以 25%为高低分组的临界值转换为二分变量。结果发现，训练的模型预测推理能力最高能获得 76.11%的查准率、65.72%的精确率、63.10%的查全率以及 65.01%的 F1 得分；预测数学成绩最高能获得 83.07%的查准率、73.70%的精确率、73.33%的查全率以及 75.57%的 F1 得分。研究结果说明，机器学习建立的区分模型具有较好的预测效果，利用 log-file 所记录的游戏过程数据可以对个体的能力进行有效预测。

关键词　电子游戏；推箱子；机器学习；推理能力；数学成绩

分类号　B849: G44

1　引言

心理测验(Psychological Testing)发展百余年来，已经被广泛应用于教育(Di Giunta et al., 2013)、企业管理(Schmidt, 2002)、临床医疗(Judd, Schettler, & Rush, 2016)等各个领域，并发挥重要作用。但传统心理测验具有的一些不足之处使其在应用上面临挑战。首先，传统心理测验由于采取明显的测验形式，容易令个体产生测验焦虑，而焦虑会显著降低受测者的外在表现，使个体真实的能力水平被低估(Cassady & Johnson, 2002)。已有研究表明，对测验成绩的焦虑会影响个体的注意(Keogh & French, 2001)、降低工作记忆表现(Ikeda, Iwanaga, & Seiwa, 1996)、降低智力测验表现并使得问题解决更加艰难(Cassady & Johnson, 2002)。更为严重的是，这种由焦虑引起的不良测验成绩会反过来作用于个体自尊，并直接导致恐惧、防御性与逃避性行为(Hembree, 1988)，对个体的长期发展十分不利。其次，某些经典的心理测验已经被使用了很多年，可能会导致测验的曝光效应，从而无法评估个体的真实水平(Bors

---

收稿日期：2017-08-10

* 北京市教育科学“十二五”规划青年专项课题(CBA15048)资助。

通信作者：黎坚, E-mail: jianli@bnu.edu.cn

& Vigneau, 2003)。研究表明，测验的频繁使用容易被人们破解所谓的"解题秘诀"，个体可能事先获得题目和答案信息而使分数提高(Neisser, 1997)，或通过不断的备考和练习来提高测评结果(Bors & Vigneau, 2003)，但这并不能反映出受测者的真实水平，反而会干扰真实的测量结果，对测评的信效度造成污染(Neisser, 1997; Hausknecht, Halpert, Di Paolo, & Moriarty Gerrard, 2007)。

鉴于传统心理测验存在的问题，研究者开始考虑采用其他形式和工具对个体能力和人格特质进行评估。近年来，随着计算机网络的发展和电子游戏的普及，基于游戏的评估方式渐渐流行起来。

基于游戏的评估(game-based assessment, GBA)是指：通过游戏(game)或者游戏化的活动(game-like activities)，来对某一对象进行评估(Heinzen, Landrum, Gurung, & Dunn, 2015)。从心理学的角度来说，即采用游戏的方式，对一个人的能力、人格等心理特性和行为进行量化评估。这里所说的游戏是指所有一般意义上的游戏，既包括生活中诸如下棋、捉迷藏等真实游戏，也包括模拟经营、伦敦塔、连连看等各类平台上的电子游戏(video game)。基于游戏的评估方法与传统的心理测验相比具有诸多优点。首先，将心理评估游戏化意味着我们可以跳出传统纸笔测验及其变式，创造出丰富而多样化的测验形式和内容，例如：视觉运动技能(visuomotor skill)的测量已经可以通过手指与触屏的游戏式互动来实现(Tenorio Delgado, Arango Uribe, Aparicio Alonso, & Rosas Díaz, 2016)。其次，游戏的可玩性、隐蔽性和仿真性等特点使得基于游戏的评估更易被人们接受，其测验情境更接近于真实生活，具有挑战性和交互式的特点，能够极大程度地提高受测者的动机和投入程度，并且还能通过降低社会称许性的影响(Heinzen et al., 2015)，在一定程度上避免传统心理测验中的测验焦虑和测验曝光问题。此外，全新的测验形式意味着可以用不同以往的方式来收集数据。例如以电子游戏作为心理测评的载体，可以通过计算机后台记录玩家的游戏任务操作过程(DiCerbo & Behrens, 2012)。对这些过程信息加以正确利用，能够让研究者对个体的能力和特质有一个更为全面的了解，也更有利于心理测评的效度提升，使得心理学研究的重点从"结果是什么"转变为分析"怎样产生结果" (Greiff, Wüstenberg, & Avvisati, 2015)。目前，基于游戏的评估得到了研究者的重视，如 Sonnleitner 和 Köstering 等人分别利用基因实验室和伦敦塔游戏任务对个体的复杂问题解决和计划性等认知能力进行了评估(Sonnleitner et al., 2012; Köstering et al., 2015); Ventura 和 Baumert 等人分别利用沙盒游戏和独裁者博弈游戏任务来评估个体的坚持性、公平性和利他性等人格特质(Ventura & Shute, 2013; Baumert, Schlösser, & Schmitt, 2014)。

然而，目前基于游戏的评估在应用中也存在一些问题，还没有在数据分析和计分逻辑中

充分利用游戏的过程信息，仅使用少数指标对某变量进行评估。例如 Li, Zhang, Du, Zhu 和 Li (2015)通过推箱子游戏测量被试的元认知计划，但仅记录并使用了第一步时间与总时间的比值，以此作为评估指标; Berg 和 Byrd (2002)在使用伦敦塔游戏测量计划性时，仅使用第一步移动前的时间作为评估指标。在上述研究中，大量过程性信息被浪费，而这些信息很可能隐藏着可以衡量个体特质或能力的指标，如每一步思考的时间、动作的回溯等(Greiff et al., 2015)，因而无法对个体特质或能力进行全面完备的评估。

鉴于基于游戏的评估在应用中所存在的问题，心理测评领域亟需一种有效的数据分析方法来处理这些问题，以实现对个体能力和特质更为准确的评估。随着人工智能研究的兴起和发展，机器学习已经成为一门集挑战性、实用性、价值性于一体的热门学科，尤其是谷歌 AlphaGo 在人机大战中获胜的奇迹使得机器学习成为备受瞩目和极富发展前景的领域。目前，已经有研究者将机器学习引入心理学领域，并应用于人格预测(Wu, Kosinski, & Stillwell, 2015)。此外, Zhang, Song, Cui, Liu 和 Zhu (2016)根据人类步态可作为情绪识别的依据这一原理，设计了内置加速度传感器的智能手环，用于测量被试的情绪状态。研究首先对 123 名被试进行了情绪操纵和原始数据收集，之后进行数据预处理和特征提取，然后通过机器学习的方法对数据进行处理和建模，结果发现机器学习的不同算法均能够较为准确地识别情绪(快乐、中性、愤怒)，尤其是 LibSVM 算法在区分中性和愤怒情绪上的准确度高达 91.3%; 在进行三种情绪状态的区分上，也达到了 81.2%的准确率。

需要指出的是，机器学习算法通常需要使用到大体量的数据，计算机过程数据追踪技术——log-file 为此提供了实现可能。所谓 log-file 是指包含受测者所有活动足迹的日志文件，其优点是能够全面、实时地记录全部数据，信息量丰富，有利于后期的数据挖掘和分析(Moharil et al., 2014)。这一技术不仅实现了对研究变量的在线测量，而且避免了传统方法中数据收集工作耗时费力的缺陷，可在大规模施测的同时无干扰地记录被试的真实行为。近年来，采用基于计算机网络平台的游戏 log-file 方式对变量进行研究已经成为心理学领域的新趋势，如 Sonnleitner 等人(2012)开发的游戏任务，会在计算机后台生成 log-file 文件，但遗憾的是他们并没有充分利用 log-file 信息，只是选取了少数几个指标用于评估受测者的复杂问题解决能力。这其中的主要原因在于数据挖掘本身的困难。log-file 数据通常变量众多、数量巨大，从这些数据中找出有意义的信息和关系是比较困难的(Csapó, Ainley, Bennett, Latour, & Law, 2012)。也正因为如此，机器学习技术可以在这类研究中体现出独特的优势，能够通过充分利用 log-file 的信息，建立较为复杂的模型，实现更为准确的预测。

基于上述分析，本研究拟采用游戏任务，并结合 log-file 技术和机器学习技术，尝试对个

体的能力倾向(aptitude)和学业成就(achievement)两类能力进行预测，在操作层面将以抽象推理能力和数学学习成绩作为研究变量。之所以选择这两类能力，是因为能力倾向和学业成绩均是学生求学生涯中的重要变量，前者能够反映出个体在广泛的信息加工任务中的潜能，后者则是学习结果的主要表现，对学生自我概念的建构、思维能力的发展以及社会适应的培养等方面都具有显著的影响(Duncan et al., 2007)。研究拟选取“推箱子”作为游戏任务。在该游戏中，受测者需要调用其认知和元认知加工，持续对箱子的位置和小人的移动路线进行思考，涉及到大量的抽象推理、空间想象与操作，以及数字运算工作，因此，受测者在游戏过程中的表现很有可能能够反映其推理能力和数学学业成就的水平。由于从游戏 log-file 中收集到的原始数据并不规范，不能直接作为特征加入模型中，因此需要研究者进行特征提取，从而生成有可能反映受测者能力的特征。以往关于推箱子的研究涉及到认知能力、元认知能力、认知效率三种指标(张博，黎坚，徐楚，李一茗，2014)。本研究在选取特征时虽然也将从这三个方面进行考虑，但会选取更多特征，通过增加特征数量使模型的预测率更高。例如在认知能力方面，拟选取完成箱子比例、关卡是否成功等特征；在元认知能力方面，拟选取第一步计划时间与总时间的比值、第一步计划时间与平均执行时间的比值，以及二者的对数等特征；在认知效率方面，拟选取思考步数、与最优步数之差等特征。此外，考虑到受测者在“成功”的关卡与“失败”的关卡中可能存在不同的表现，因此在特征选取时对两种通关情况也进行了区分。

2　方法

研究包含 4 个步骤：数据采集、特征提取、数据预处理和模型训练。

2.1　数据采集

2.1.1　被试

首都师范大学第一附属中学的 395 名初一和初二学生参加了研究，有效被试为 360 名，其中女生 172 名，男生 188 名。初一学生平均年龄为 13.2 岁，最小为 11.9 岁，最大为 15.1 岁；初二学生平均年龄为 14.6 岁，最小为 14.1 岁，最大为 16.3 岁。

2.1.2　推箱子游戏

推箱子(Sokoban)是一款经典益智游戏，由一个小人和至少 1 个箱子组成，玩家需要操纵小人上下左右移动，将箱子推入目标位置。经过研究者对程序的改编，计算机后台可以自动生成 log-file 文件，记录受测者每一步的潜伏期与按键动作，例如时间节点、小人移动的方向、推动的箱子、完成步数等。游戏一共 23 题，所有题目都不允许悔棋。前 3 题为练习题，用于

帮助受测者熟悉键盘操作及游戏规则。在练习题阶段，答错可重来，直到全部通过方可进入正式游戏。第 4~23 题为正式题目，每题限时 2min。与练习阶段不同，正式题目只有一次机会，若一次未通过，则只能放弃该题。正式题目中的箱子个数在 1~3 之间，每题均有唯一最优解。游戏界面如图 1 所示。

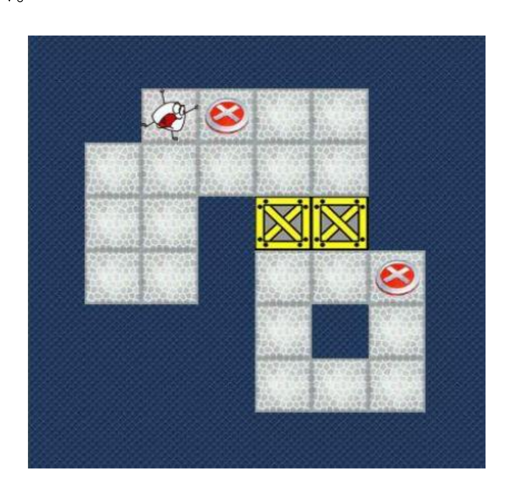


图 1 推箱子游戏界面截图

2.1.3 测量工具

瑞文标准推理测验(Standard Progressive Matrices, SPM; Raven, 1989)：用于测量一般智力中的抽象推理能力，共计 60 个条目，每答对一题计一分，满分 60 分，总分即为推理能力得分。在本研究的受测者中，平均分为 46.2 分，标准差为 8.5。

数学成绩：研究者获得了学生的三次数学测验的成绩(期中、月考、期末)，每次数学测验均为年级内统一施测，因此在年级内具有可比性。将三次数学成绩取平均值以获得对学生数学能力较为准确的估计。本研究中，数学成绩的平均分为 64.9 分，标准差为 19.9。

本研究属于机器学习中的监督学习(supervised learning)类别，而监督学习分为两种，一是利用分类算法预测标称型数据，二是利用回归算法预测连续型数据。对于本研究而言，推理能力和数学成绩是连续型数据，理论上应使用回归算法进行预测，但是回归算法对特征数量和样本量的要求较高，其创建的模型需要拟合所有的样本点，当数据拥有众多特征且特征之间的关系十分复杂时，构建全局模型难以实现(Harrington, 2013)。囿于现实因素，研究无法获取更多受测者，且众多特征间的关系较为复杂，坚持使用回归算法将无法达到良好的预测效果。因此，作为一项尝试性研究，本研究在数据分析中将两个结果变量转化为标称型数据，具体做法是：将瑞文推理测验得分在前 25%的学生记为 1，后 25%得分的学生记为 0，构造瑞文得分的二分变量作为最终结果变量(即机器学习算法的标签)，其他学生的数据不予使用，最终有效的受测者是 180 名；将数学成绩按照年级分层，将每个年级数学成绩排名前 25%

的学生标记为1，后25%的学生标记为0，构造数学成绩的二分变量作为最终结果变量，其余学生的数据不予使用，最终有效受测者也为180名。

2.1.4 测试流程

所有学生均以班级为单位在计算机教室参加测试。受测者首先在计算机上完成推箱子游戏任务，然后填写纸质版的瑞文推理测验。施测过程都由任课教师、班主任和主试一同监督，保证安静、独立作答。

2.2 特征提取

程序记录的受测者顺利通过某一关的典型数据形式如图2所示。通常，在第一步之前会有较长时间的思考，在思考完成后则有较快的按键反应，在操作过程中，偶尔会出现潜伏期。除了过程数据外，程序同样记录了受测者在本关是否通过、有几个箱子被移到了指定位置、受测者是否主动放弃本关、任务超时等信息。下面将对特征提取的具体方法进行说明。

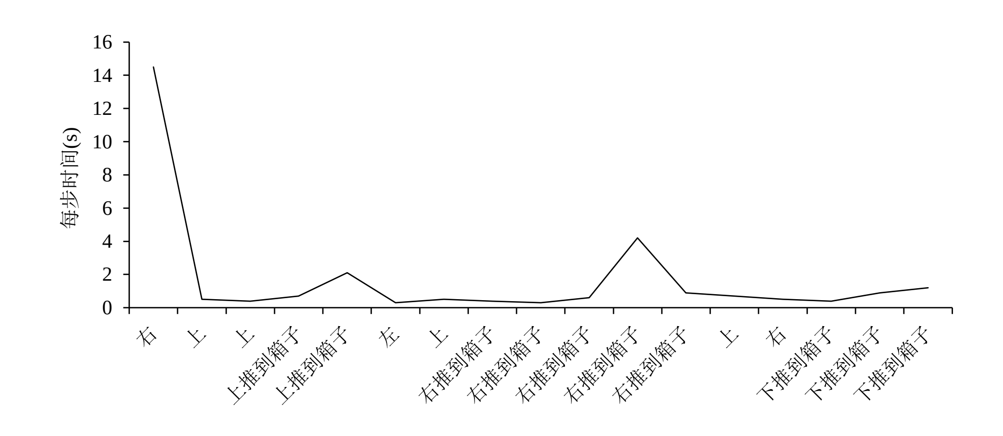


图 2 一个典型的行动过程

2.2.1 第一步所用时间

由图2可知，受测者通常会分配较长时间在第一步之前的思考中。参照文献，我们计算出受测者在每一关中第一步所用时间占总时间的比重。此外，用第一步时间除以平均执行时间(定义方法见2.2.3)以及对上述变量取其对数形式，构造出多个特征以进行较为全面的衡量。

2.2.2 执行间思考

受测者在执行过程中可能会停下来思考，反映在数据中即是：在执行过程中某一步用时较其他时间异常变高。为了反映这种波动性，计算受测者除了第一步之后各步用时的标准差，即执行间波动，同时，记录时间超过平均值一个标准差以上的步数占总步数的比例作为受测者在执行过程中的思考次数的指标。

2.2.3 平均执行时间

剔除掉包含思考的行动后(包括第一步)，将余下的行动时间求平均。这部分时间反映了受测者在无需思考的情况下执行操作的时间。

2.2.4 冗余步数

根据受测者的行动路线可以求出受测者每一步行动后的整体状态。计算出其路径中重复的状态(即在两个状态中，小人和所有箱子的位置完全相同)占最优路径中总状态数的比例。其中，对于一个状态重复多次的情况只记为 1，以避免玩家在两个状态间“来回踱步”产生误差。

2.2.5 与最优路径重合比例

通过广度优先搜索算法，可以求出每一关的最优路径，并计算受测者的路径与最优路径重合的比例。具体而言，计算受测者路径的状态集合与最优路径状态集合的交集，计算交集占最优路径状态集的比例，同样排除掉“来回踱步”的情况。

2.2.6 与最优路径相差步数

受测者步数与最优步数的差异，也作为一个指标加入模型中。

2.2.7 完成箱子的比例

所有关卡中，程序都会报告受测者完成的箱子占总箱子的比例，把它作为一个特征加入到模型中。

2.2.8 每题是否成功、放弃

每道题有三种状态，成功通过、放弃、超时，用两个二分变量来刻画这三种情况，并作为特征加入模型中，即对于“是否成功”这一特征来说，成功记为 1，放弃记为 0，超时记为 0；对于“是否放弃”这一特征来说，成功记为 0，放弃记为 1，超时记为 0。

2.3 数据预处理

原始数据以每一关为一个观测值，但在训练模型时需要以受测者为单位的观测值。直接将长型数据转换为宽型数据，即对于每个受测者每一关都提取出上述特征，总共特征数是单关特征数的 20 倍，这会存在以下问题：一，由于样本规模不大，特征过多不利于模型训练，容易过拟合；二，同一个指标在成功和失败两种状态下可能有不同的意义，以“与最优路径相差步数”这一特征为例，失败的状态下，受测者的步数通常会低于最优路径步数，数字越大表明受测者越坚持，而在成功的状态下，数字越大则反应受测者的步数偏离最优路径越远。为了克服以上问题，将上述特征均以成功与否划分为两组，即对于以上每个特征，都构造出两类：一类用于描述该特征在成功的关卡中对推理能力和数学成绩的预测能力，另一类描述在失败的关卡中对二者的预测能力。对于是否成功、是否放弃两个二分变量，直接在各组间

求平均容易忽略掉每道题的难度信息，故使用因子分析从其中提取出两个因子作为特征用以训练模型。因此，模型中共放入 23 个特征用于训练，特征的描述统计结果见表 1。

表 1　特征的描述统计结果

特征	平均值	标准差	最小值	最大值
失败组				
第一步用时/平均执行时间	22.71	24.26	2.52	198.34
ln (第一步用时/平均执行时间)	2.31	0.82	0.81	4.97
完成箱子的比例	0.33	0.08	0.00	0.57
第一步用时/总时间	0.22	0.12	0.04	0.76
ln (第一步用时/总时间)	−1.92	0.60	−3.31	−0.29
思考步数占比	−2.39	0.23	−3.04	−1.69
平均执行时间	0.64	0.15	0.37	1.33
执行间波动	2.15	1.20	0.35	10.52
重复步数占比	0.07	0.03	0.00	0.20
与最优步数相差	−5.75	9.45	−23.36	65.78
与最优路径重合步数占比	0.17	0.04	0.04	0.32
成功组				
第一步用时/平均执行时间	24.36	23.81	2.65	168.97
ln (第一步用时/平均执行时间)	2.49	0.78	0.92	4.95
第一步用时/总时间	0.25	0.14	0.04	0.77
ln (第一步用时/总时间)	−1.77	0.61	−3.18	−0.27
思考步数占比	−2.61	0.27	−3.53	−1.64
平均执行时间	0.48	0.11	0.33	1.18
执行间波动	1.17	0.76	0.20	5.43
重复步数占比	0.03	0.02	0.00	0.16
与最优步数相差	7.65	5.45	0.00	52.67
与最优路径重合步数占比	0.71	0.14	0.17	1.06

注：因子分析抽取出的两个因子不具有平均数、标准差等描述性统计含义，故未放在表格中。

研究者计算了上述特征与瑞文测验成绩和数学成绩的相关，结果发现成功组和失败组的第一步用时/总时间、ln (第一步用时/总时间)、第一步用时/平均执行时间、ln (第一步用时/平均执行时间)等特征均与瑞文测验成绩和数学成绩有显著相关，相关系数在 0.19~0.46 之间。此外，数学成绩还与失败组思考步数占比、失败组完成箱子的比例显著相关，相关系数分别是 0.16 和 0.17。这些结果初步表明了本研究特征选取的有效性。

2.4　模型训练

推理能力与数学成绩的模型训练策略一致，使用基于 Python 3 的 scikit-learning 包

(Pedregosa et al., 2011)提供的随机森林模型进行训练，该算法是分类学习的常用算法。随机森林(Random Forests, RF)是决策树的集合，利用多棵树对样本进行训练和评估的分类器。该算法会重复选择随机样本，在训练集中生成多个样本集，每个样本集都会形成一棵树，最后根据生成的这些树在测试集中进行评估，投票最多的作为最终类标签(Breiman, 2001)。

研究中，首先随机划出 30%的样本作为评估集，70%的样本用于交叉验证以及超参数搜索。在 70%的样本中，使用 4 折交叉验证策略将样本再次随机分成 4 组，依次选择其中一组作为验证集，其他三组作为训练集。在训练集上训练特定参数的模型，将获得的模型在验证集中测试，计算模型得分。4 个轮次后，每组样本均有 3 次作为训练集，1 次作为验证集，将 4 组中获得的得分求平均，为对应超参数的模型在交叉验证组中的得分。

随机森林模型具有多个参数可供调整，不同的参数设置会影响模型的拟合效果。本研究关注的主要参数为：最大特征数、最大深度、最小分裂样本量、拟合器数量。最大特征数表示在寻找一个最优的分裂过程中需要考虑的特征数；最大深度表示一个决策树最大的深度，达到该深度后即停止分裂；当一个节点上的样本量低于最小分裂样本量时该节点即停止分裂；拟合器数量是指在随机森林中的树的数量。采用网格搜索的策略寻找出最优的参数，对每一种超参数组合都进行一次交叉验证，选择出交叉验证中平均得分最高的超参数组合。搜索的范围为：最大特征数由 5 至 16；最小分裂样本量由 2 至 10；最大深度由 2 至 8，拟合器数量包括 5、10、50、160 四种。因此总共需要进行 3024 次交叉验证。经过 3024 次交叉验证后，在验证集上得分最高的超参数组合即为最优参数组合。但是通过交叉验证获得的得分并不能准确评估该模型的有效性：因为超参数搜索本身也相当于拟合的过程，容易出现过拟合。因此对于该模型的评估需要在评估集上测试。首先，使用获得的最优参数，用交叉验证中涉及的所有 70%的样本对随机森林模型进行拟合，之后用拟合后的模型以及评估集中的特征对推理能力分类和数学成绩分类进行预测，用预测得到的分类与真实分类作比较，计算得分。由此得到的得分是对模型预测能力比较准确的评估。

分类模型建立后，可以用多种评估方式考察模型的预测效果，主要可以分为 4 类：真实阳性(True Positive)，即预测为阳性且实际上也为阳性；虚假阳性(Fake Positive)，即预测为阳性但实际上为阴性；真实阴性(True Negative)，即预测为阴性实际上也为阴性；虚假阴性(Fake Negative)，即预测为阴性但实际上是阳性。每种类别的表示方式如表 2 所示。

表 2　分类表现评估表

表现类型	预测为阳性	预测为阴性
实际为阳性	TP	FN
实际为阴性	FP	TN

根据上述分类可以得到最常用的指标：

(1)精确率(Accuracy)：精确率是最为简单、直接的一个指标，为正确预测的观测值数量除以总观测值数量，即(TP+TN)/(TP+TN+FP+FN);

(2)查准率(Precision, P)：查准率描述在预测为阳性的样本中，真实阳性的比例，即TP/(FP+TP);

(3)查全率(Recall, R): 查全率描述在所有实际阳性样本中，预测为阳性的比例，即TP/(TP+FN);

(4) F1：查准率与查全率两者显然是存在一些矛盾：追求查准率则会牺牲一些查全率，反之亦然。F1 则是查准率与查全率两者之间较为平衡的一个指标，公式为 F1 = 2RP/(R+P)。

在超参数搜索中，分别以 4 种计分方式为目标，即对应每个计分方式均找到一个能使其最大化的超参数组合，以满足不同的预测需要。

3　结果

通过对模型的调整可以发现一个模型中所有特征的平均重要性，特征重要性定义为：标准化后的特征减少的基尼不纯度(Tan, Steinbach, & Kumar, 2006)。图 3 显示的是数学成绩预测模型中排列前十的特征(推理能力预测模型与此类似)，可见，第一步时间与平均执行时间的比值取对数(成功组)在数学成绩预测模型中的平均重要性最高；第二重要的特征是从成功与否指标中通过因子分析提取出来的第一个因素；与最优解相差的步数、思考时间占比、执行间波动、重复步数占比等因素均对模型有一定贡献，证实了研究所提取的特征的有效性。

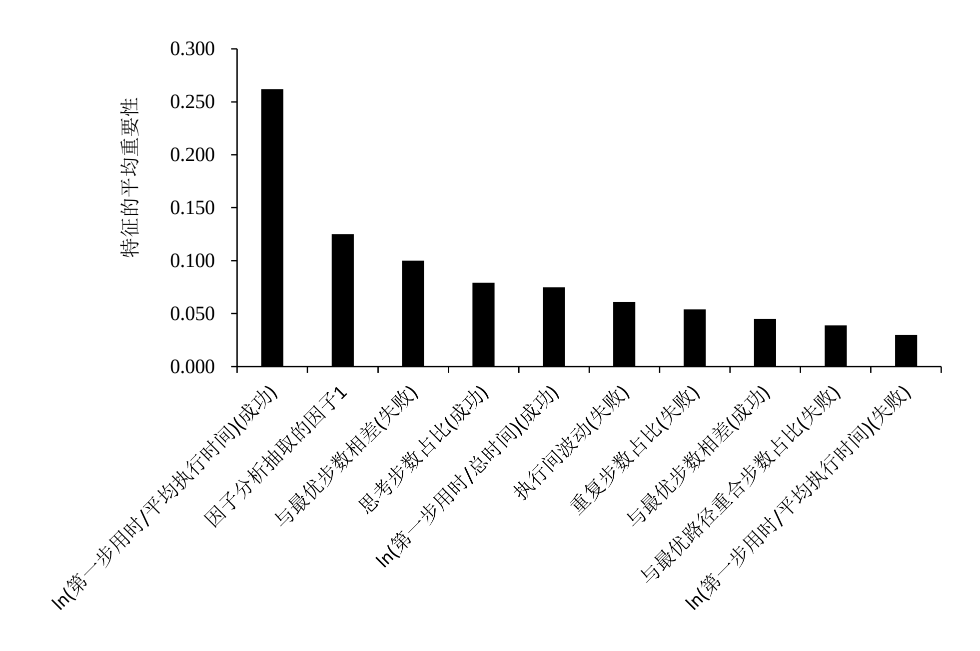


图 3 数学成绩预测模型中平均重要性排列前十位的特征

通过对上文所述的特征进行训练，所得模型在评估集上的表现如表 3 所示。第一列表示在超参数搜索时评估模型使用的标准，即最优化目标，其他四列表示相应的模型在评估集上相应指标的得分。结果表明，在超参数搜索阶段采用不同的最优化目标对结果的影响不大。用该模型预测数学成绩，能够达到 80%左右的查准率，而 F1、查全率、精确率都在 70%左右；对于推理能力的预测结果稍弱于对数学成绩的预测，平均能达到 65%左右的 F1, 75%左右的查准率, 60%左右的查全率，以及 65%左右的精确率。该结果表明机器学习建立的区分模型结果尚可接受。

表 3　模型预测结果

最优化目标	F1	查准率	查全率	精确率
推理能力				
F1 优先	68.83%	74.40%	61.19%	63.46%
查准率优先	63.72%	75.51%	59.17%	65.03%
查全率优先	65.01%	74.91%	63.10%	64.21%
精确率优先	64.22%	76.11%	59.05%	65.72%
数学成绩				
F1 优先	71.14%	79.35%	71.11%	68.02%
查准率优先	75.57%	83.07%	73.33%	73.70%
查全率优先	73.09%	81.06%	71.78%	70.62%
精确率优先	71.65%	80.19%	69.67%	69.44%

4 讨论

本研究尝试从推箱子的 log-file 数据中挖掘信息，并使用随机森林模型预测学生的推理能力分类以及数学成绩分类，模型预测的推理能力达到了约 75%的查准率以及 65%左右的精确率；预测数学成绩达到了约 80%的查准率以及 70%左右的精确率，结果说明模型预测效果尚可，表明游戏 log-file 结合机器学习能够训练出较好的区分模型用以预测个体的抽象推理能力和学业成就。

4.1 过程性数据的利用

研究通过从 log-file 的过程性数据中提取多个特征，一方面部分支持了以往研究的结果，另一方面也完善了前人研究中特征利用不充分的不足。

首先，本研究发现，计划性指标(即第一步所用时间占全部时间的比值)在推理能力和数学成绩的区分模型中均占有重要地位，且成功组的比值大于失败组，说明事先计划对于开展逻辑推理类任务至关重要。无论是在完成图形推理题目，还是在回答数学试题方面，对题目或任务的计划会在一定程度上决定最终的结果，这与以往研究相一致(Li et al., 2015)。

其次，以往使用推箱子的研究没有区分受测者在题目上成功和失败两种情况，而同一特征在两种情况下可能代表不同的含义，这样做会浪费特征本身具有的价值。本研究在获取大量过程性数据的基础上，对特征在成功或失败情况下进行了分解，以便充分利用每个特征。结果发现，同一特征在不同情况下的贡献确实存在差异。例如，在数学成绩预测模型中，ln (第一步用时/平均执行时间)这个特征在成功和失败情况下均对模型有贡献，但前者的贡献要大于后者；在模型贡献排名前十的特征中，失败情况下有较强预测效果的是“与最优步数差”、“执行间波动”、“重复步数占比”、“与最优路径重合比例”等特征；而成功情况下有较强预测效果的是“思考步数占比”、“与最优步数差”、“ln (第一步时间/总时间)”等特征。

4.2 机器学习算法的预测效果

对于推理能力和数学成绩这两个结果变量来说，使用随机森林所建立的区分模型在预测效果上存在一定差异。具体而言，数学成绩的区分模型能够达到 83.07%的查准率，而对于推理能力的区分模型，最高能达到 76.11%的查准率。该结果一方面说明本研究中的随机森林模型在查准率指标上达到了较为一致的最优预测效果，另一方面说明从推箱子任务中提取的现有特征更适用于预测数学成绩。这可能是因为，瑞文推理测验以测量抽象推理能力为主，相对来说对能力的要求比较单一；但数学考题涉及的认知和元认知能力更加广泛，包括对数量关系的梳理、空间图形的表征、使用策略的选择等，这与推箱子游戏的问题解决过程更为

相似。

在模型训练中，使用不同的最优化目标会对模型预测结果产生一定影响。对于推理能力，采用精确率优先条件时会得到最高的预测查准率；对于数学成绩，采用查准率优先条件时会得到最高的预测查准率。说明对于不同的结果变量，不同的最优化指标会产生不同的结果，基本不存在可以同时适用于多个变量的模型。在不同的模型中，参数、最优化指标会有所不同，最后的结果也会呈现一定的差异。机器学习的好处就是可以通过不断的调配参数来获取对结果变量最为有效的预测性，选择具有最优预测效果的模型。

4.3 游戏 log-file 和计算机技术在心理测量中的应用

随着信息技术的发展，计算机过程数据分析技术(computer logfiles analysis)在心理测量领域的地位不断提升。计算机过程数据分析技术是指通过追踪、分析受测者在计算机上完成任务过程中的操作行为、操作时间等信息来提取测量指标(Veenman, Bavelaar, De Wolf, & van Haaren, 2014)。该技术相较于传统的技术有诸多显而易见的优点。首先，传统的测量方法基于被试的自我报告，或者基于主试的对于被试行为的编码，而过程数据的分析依赖于客观指标，能够更好地达到标准化(Veenman, Wilhelm, & Beishuizen, 2004)；其次，传统的测量方法，尤其是在测量认知能力时，会有较强的侵入性：如观察、出声思考技术(Pressley & Afflerbach, 1995)等，在这些条件下受测者完成任务的能力一定程度上会受到影响；而过程数据分析则不具有侵入性，受测者完成任务的过程中不会受到打扰，能够在最自然地状态下表现(Veenman et al., 2014)；最后，过程分析技术成本较低，可以对多个受测者同时施测，数据分析通过计算机自动化完成，因此相较于传统的测量技术如出声思考法以及眼动追踪技术(Kinnunen & Vauras, 1995)等更加省时。由于大量的过程数据无法用传统统计方法进行分析，因而需要机器学习算法的引入和使用。这些优点使得游戏 log-file、计算机过程数据分析技术、机器学习将在未来得到更多的发挥空间。

4.4 研究意义与局限性

研究利用推箱子游戏获取的 log-file 信息，并使用机器学习算法，建立了较为有效的区分模型，初步实现了对学生的能力倾向和学业成就的预测，结果可以为其他心理学和教育学的测评工作提供借鉴。

同时，本研究也存在一定局限性。首先，特征提取和模型建立过程的样本仅包含测验得分排在前 25%与后 25%的受测者，两个群体间能力差异较大，在这种情况下，80%的查准度并不完美，并且如果未来的研究目标转化为对连续变量进行预测时，预测的难度会进一步提高；其次，虽然目前基于推箱子游戏的过程数据建立的模型可以较好地预测个体在瑞文推理

测验和数学测验上是否成功，但实际上我们对推箱子这一游戏任务本身到底涉及到哪些认知加工过程并不完全清楚，因此并不能直接用模型预测的分数代替瑞文推理测验成绩或数学测验成绩。在之后的研究中，可以考虑从以下几个方面进行改进。第一，增大样本量。由于本研究在模型训练中使用的样本量低于 200，为了避免过拟合，必须减少特征数量，因而限制了模型的拟合能力。未来研究如果能够收集到足够大的样本量，则可以使用更多的特征，从而提高模型的预测能力。第二，增大受测者在所预测能力上的异质性。本研究采用的样本均来自于同一所中学，受测者的抽象推理能力相对来说具有较强的同质性，这将增大模型预测的难度。未来研究如果能在不同的地区(城市、农村)、不同水平的学校(重点、普通)取样，也可能会提高模型的预测能力。第三，提取更多的特征并尝试其他算法。相对于以往研究，本研究从推箱子游戏的过程数据中提取了更多样化的指标，包括重复状态比例、思考次数等，但 log-file 中仍有很多信息可以被挖掘。未来的研究可以尝试提取和检验更多的特征，并可考虑采用机器学习中的回归算法对能力倾向和学业成就进行预测。第四，相对于瑞文推理测验任务，推箱子游戏任务所包含的认知加工过程更加多样化，更具动态性，未来研究可以考虑结合眼动或者 fMRI 技术对这一游戏任务所涉及的心理加工过程进行更深入的考察，从而指导研究者构造出更多有实质意义的特征，为能力的个体差异的预测提供更有价值的预测源，甚至通过机器学习建模来替代某些传统能力测量工具。

5　结论

(1)可以利用游戏 log-file 收集受测者在任务中的过程性数据，并利用机器学习算法进行分析。

(2)机器学习建立的区分模型对数学成绩和推理能力具有较好的预测效果，利用电脑所记录的游戏过程数据可以对个体的能力进行有效预测。

参考文献

Baumert, A., Schlösser, T., & Schmitt, M. (2014). Economic games: A performance-based assessment of fairness and altruism. *European Journal of Psychological Assessment*, *30*(3), 178–192.

Berg, W. K., & Byrd, D. L. (2002). The Tower of London spatial problem-solving task: Enhancing clinical and research implementation. *Journal of Clinical and Experimental Neuropsychology, 24*(5), 586–604.

Bors, D. A., & Vigneau, F. (2003). The effect of practice on Raven's Advanced Progressive Matrices. *Learning and Individual Differences, 13*(4), 291–312.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology, 27*(2), 270–295.

Csapó, B., Ainley, J., Bennett, R. E., Latour, T., & Law, N. (2012). Technological issues for computer-based assessment. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 143–230). Dordrecht: Springer.

DiCerbo, K. E., & Behrens, J. T. (2012). Implications of the digital ocean on current and future assessment. In R. W. Lissitz & H. Jiao (Eds.), *Computers and their impact on state assessments: Recent history and predictions for the future* (pp. 273–306). Charlotte, NC: Information Age Publishing.

Di Giunta, L., Alessandri, G., Gerbino, M., Kanacri, P. L., Zuffiano, A., & Caprara, G. V. (2013). The determinants of scholastic achievement: The contribution of personality traits, self-esteem, and academic self-efficacy. *Learning and Individual Differences, 27*, 102–108.

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P.,... Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*(6), 1428–1446.

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education, 91*, 92–105.

Harrington, P. (2013). *Machine learning in action* (R. Li, P. Li, Y. D. Qu, & B. Wang, Trans.). Beijing, China: Posts & Telecom Press.

[Harrington, P. (2013). 机器学习实战 (李锐, 李鹏, 曲亚东, 王斌 译). 北京: 人民邮电出版社.]

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Moriarty Gerrard, M. O. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92*(2), 373–385.

Heinzen, T. E., Landrum, R. E., Gurung, R. A. R., & Dunn, D. S. (2015). Game-based assessment: The mash-up we've been waiting for. In T. Reiners & L. C. Wood (Eds.), *Gamification in education and business* (pp. 201–217). Switzerland: Springer International Publishing.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research, 58*(1), 47–77.

Ikeda, M., Iwanaga, M., & Seiwa, H. (1996). Test anxiety and working memory system. *Perceptual and Motor Skills, 82*(3), 1223–1231.

Judd, L. L., Schettler, P. J., & Rush, A. J. (2016). A brief clinical tool to estimate individual patients' risk of depressive relapse following remission: Proof of concept. *American Journal of Psychiatry, 173*(11), 1140–1146.

Keogh, E., & French, C. C. (2001). Test anxiety, evaluative stress, and susceptibility to distraction from threat. *European Journal of Personality, 15*(2), 123–141.

Kinnunen, R., & Vauras, M. (1995). Comprehension monitoring and the level of comprehension in high-and low-achieving primary school children's reading. *Learning and Instruction, 5*(2), 143–165.

Köstering, L., Schmidt, C. S. M., Egger, K., Amtage, F., Peter, J., Klöppel, S.,... Kaller, C. P. (2015). Assessment of planning performance in clinical samples: Reliability and validity of the Tower of London task (TOL-F). *Neuropsychologia, 75*, 646–655.

Li, J., Zhang, B., Du, H., Zhu, Z., & Li, Y. M. (2015). Metacognitive planning: Development and validation of an online measure. *Psychological Assessment, 27*(1), 260–271.

Moharil, B., Gokhale, C., Ghadge, V., Tambvekar, P., Pundlik, S., & Rai, G. (2014). Real time generalized log file management and analysis using pattern matching and dynamic clustering. *International Journal of Computer Applications, 91*(16), 1–6.

Neisser, U. (1997). Rising scores on intelligence tests: Test scores are certainly going up all over the world, but whether intelligence itself has risen remains controversial. *American Scientist, 85*(5), 440–447.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O.,... Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research, 12*, 2825–2830.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, N.J.: Erlbaum.

Raven, J. (1989). The raven progressive matrices: A review of national norming studies and ethnic and socioeconomic variation within the united-states. *Journal of Educational Measurement, 26*(1), 1–16.

Schmidt, F. L. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human Performance, 15*(1–2), 187–210.

Sonnleitner, P., Brunner, M., Greiff, S., Funke, J., Keller, U., Martin, R.,... Latour, T. (2012). The *Genetics Lab*: Acceptance and psychometric characteristics of a computer-based microworld assessing complex problem solving. *Psychological Test and Assessment Modeling, 54*(1), 54–72.

Tan, P. N., Steinbach, M., & Kumar, V. (2006). *Introduction to data mining*. India: Pearson Education.

Tenorio Delgado, M., Arango Uribe, P., Aparicio Alonso, A., & Rosas Díaz, R. (2016). TENI: A comprehensive

battery for cognitive assessment based on games and technology. *Child Neuropsychology, 22*(3), 276–291.

Veenman, M. V. J., Wilhelm, P., & Beishuizen, J. J. (2004). The relation between intellectual and metacognitive skills from a developmental perspective. *Learning and Instruction, 14*(1), 89–109.

Veenman, M. V. J., Bavelaar, L., De Wolf, L., & van Haaren, M. G. P. (2014). The on-line assessment of metacognitive skills in a computerized learning environment. *Learning and Individual Differences, 29*, 123–130.

Ventura, M., & Shute, V. (2013). The validity of a game-based assessment of persistence. *Computers in Human Behavior, 29*(6), 2568–2572.

Wu, Y. Y., Kosinski, M., & Stillwell, D. (2015). Computer-based personality judgments are more accurate than those made by humans. *Proceedings of the National Academy of Sciences of the United States of America, 112*(4), 1036–1040.

Zhang, B., Li, J., Xu, C., & Li, Y. M. (2014). The developmental differences of problem solving ability between intellectually-gifted and intellectually-average children aged from 11-14 years old. *Acta Psychologica Sinica, 46*, 1823–1834.

[张博，黎坚，徐楚，李一茗. (2014). 11~14 岁超常儿童与普通儿童问题解决能力的发展比较. *心理学报*, *46*, 1823–1834.]

Zhang, Z., Song, Y. F., Cui, L. Q., Liu, X. Q., & Zhu, T. S. (2016). Emotion recognition based on customized smart bracelet with built-in accelerometer. *PeerJ, 4*, e2258.

Using game log-file to predict students' reasoning ability and mathematical achievement: An application of machine learning

SUN Xin[1]; LI Jian[1,2]; FU Zhiyu[1]

([1] *Faculty of Psychology, Beijing Normal University;* [2] *Beijing Key Lab of Applied Experimental Psychology, Beijing* 100875, *China*)

Abstract

With the development of the progress of information technology, the deficiency of traditional psychological testing is becoming more obvious, such as test anxiety and test exposure. Some researchers have begun to test individuals using game-based assessment, which has many advantages, such as increasing the motivation and input level of the participants, and providing the

possibility for the implementation of log-file technology. However, the current data analysis and scoring logic ignore substantial information of process, and thus cannot accurately assess individual characteristics and abilities. The advantages of machine learning in data analysis provide a new direction. The machine learning algorithm can analyze the log-file data by building a complex model.

The present study attempted to use game-based assessment combining game log-file and machine learning techniques to predict participants' ability: reasoning ability and mathematical achievement. Participants were 360 first and second grade students from a middle school in Beijing; predictive variables were a series of features extracted from the game log-file, outcome variables were dichotomous variables calculated from Raven test and mathematics achievement, which took 25th and 75th percentile as the cutoff line. In the model training, the random forest algorithm was selected, 70% samples were randomly selected for cross validation and hyper parametric search, and then the prediction was carried out on the other 30% of samples.

Results showed that the logarithm of the ratio of the first step time to the average execution time was the highest features of average importance ratio, and the number of steps that are different from the optimal solution, thinking time ratio, execution between fluctuation, proportion of repeat steps all contributed to the mathematical achievement prediction model; reasoning ability prediction model was similar. With these important features, it could be found that the reasoning ability prediction model had 76.11% precision, 65.72% accuracy, 63.10% recall and 65.01% F1 scores; the mathematical achievement prediction model had 83.07% precision, 73.70% accuracy, 73.33% recall and 75.57% F1 score.

The finding of the present study showed that the random forest model had acceptable predictive effect when predicting reasoning ability and mathematics achievement classification based on the game log-file, with 75% precision of reasoning and 80% precision of math. In conclusion, the research provides a new method to predict the cognitive ability and academic achievement of the students; the game log-file combined with machine learning can establish an effective discrimination model. This result can provide some reference and direction for the development of educational psychological assessment.

Key words video game; Sokoban; machine learning; reasoning ability; mathematical achievement.